A classical measure of evidence for general null hypotheses

Alexandre Galvão Patriota

Departamento de Estatística, IME, Universidade de São Paulo
Rua do Matão, 1010, São Paulo/SP, 05508-090, Brazil
email: `patriota@ime.usp.br`

Abstract

In science, the most widespread statistical quantities are perhaps p-values. A general advice is to reject the null hypothesis H_0 if the corresponding p-value is sufficiently small (usually smaller than 0.05). Many criticisms regarding p-values have arisen in the scientific literature. The main raised issue is that p-values are not measures of evidence (support measures) over the parametric space Θ. Here, we propose a *frequentist* **measure of evidence** for very general null hypotheses that satisfies logical requirements on the subsets of Θ that are not met by p-values (e.g., it is a possibilistic measure). We study the proposed measure in the light of the abstract belief calculus formalism and we conclude that it can be used to establish objective states of belief on the subsets of Θ. Based on its properties, we strongly recommend this measure as an additional summary of significance tests. At the end of the paper we give a short listing of possible open problems.

Keywords: Abstract belief calculus, evidence measure, likelihood-based confidence, nested hypothesis, p-value, possibilistic measure, significance test

1 Introduction

Tests of significance are subjects of intense debate and discussion among many statisticians (Kempthorne, 1976; Cox, 1977; Berger and Sellke, 1987; Aitkin, 1991; Schervish, 1996; Royall, 1997; Mayo and Cox, 2006), scientists in general (Dubois and Prade, 1990; Darwiche and Ginsberg, 1992; Friedman and Halpern, 1996; Wagenmakers, 2007) and philosophers of science (Stern, 2003; Mayo, 2004). In this paper, we discuss some limitations of p-values (which is a well-explored territory) and we propose an alternative measure to establish objective states of belief on the subsets of the full parametric space Θ. Currently, many scholars have been studying the controversies and limitations of p-values (see, for instance, Mayo and Cox, 2006; Mayo and Spanos, 2006; Wagenmakers, 2007; Pereira *et al.*, 2008; Rice, 2010; Grendár,

1

2012; Diniz *et al.*, 2012) and others have proposed some alternatives (Zang, 2009; Bickel, 2010). In this paper, besides proposing a classical (read frequentist) measure of evidence, we also provide a connection with the abstract belief calculus (ABC) proposed by Darwiche and Ginsberg (1992), which certifies the status of "objective state of belief" for our proposal, see Section 4 for specific details.

A procedure that measures the consistency of an observed data x (the capital letter X denotes the random quantity) with a null hypothesis H_0 is known as a significance test (Kempthorne, 1976; Cox, 1977). According to Mayo and Cox (2006), to do this in the frequentist paradigm, we may find a function $t = t(x)$ called test statistic such that: (1) the larger the value of t the more inconsistent are the data with H_0 and (2) the random variable $T = t(X)$ has a known probability distribution when H_0 is true (at least asymptotically). The p-value related to the statistic T (the observed level of significance Cox, 1977) is then computed as

$$p = P(T > t; \text{under } H_0),$$

which is regarded as a measure of concordance with H_0 (Mayo and Cox, 2006). A p-value is a probability of an unobserved T be, at least, as extreme as the observed t when H_0 is true. This means that small values of p indicate a discordance of the data probabilistic model from that specified in H_0. It is of large agreement to set in advance a threshold value α to reject H_0 if and only if $p \leq \alpha$. Pereira and Wechsler (1993) point out some problems when the statistic T does not consider the alternative hypothesis. Schervish (1996) had argued that p-values as measures of evidence for hypotheses has serious logical flaws. Berger and Sellke (1987) argue that p-values can be highly misleading measures of the evidence provided by the data against the null hypothesis. In this paper, we present two examples where conflicting conclusions arise if p-values are used to take decisions regarding the inadequacy of a hypothesis.Then, we propose a measure of evidence for general null hypotheses that is free of those conflicts and has some important philosophical implications in the frequentist paradigm which will be detailed in future works.

Here, the null hypothesis is defined in a parametric context, let $\theta \in \Theta \subseteq \mathbb{R}^k$ be the model parameter, the null hypothesis is defined as $H_0 : \theta \in \Theta_0$, i.e., if H_0 is true, then the true unknown value of the parameter vector lies in the subset $\Theta_0 \subseteq \Theta$. The parametric space of the alternative hypothesis will always be the complement of Θ_0 with respect to Θ, therefore, if the alternative hypothesis is specified the full parametric space has to take it into account. A hypothesis test attempts to reduce a family of possible measures that governs the data behavior, say $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$, to a more restricted one, say $\mathcal{P}_0 = \{\mathbb{P}_\theta : \theta \in \Theta_0\}$. There are many ways to find a test statistic, it essentially depends on the topologies of Θ_0

and Θ. When Θ_0 and its complement have one element each, the Neyman-Pearson Lemma provides the most powerful test (which is the likelihood ratio statistic) for any pre-fixed significance value. Naturally, we can use this statistic to compute a p-value. For the general case, the generalization of likelihood ratio statistic (which will be called only by likelihood ratio statistic) is given by

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta; x)}{\sup_{\theta \in \Theta} L(\theta, x)}$$

where $L(\theta, x)$ is the likelihood function. Observe that the likelihood ratio statistic does take into account the alternative hypothesis (since $\Theta = \Theta_0 \cup \Theta_1$ with Θ_1 being the parametric space defined in the alternative hypothesis). The reader should notice that the likelihood ratio statistic is a generalization for uniformly most powerful tests (see Birkes, 1990) under general hypothesis testing. For general linear hypothesis, $H_0 : C\theta = d$, we can also resort to a Wald-type statistic

$$W(x) = (C\widehat{\theta} - d)^{\top} [CAC^{\top}]^{-1} (C\widehat{\theta} - d)$$

with $\widehat{\theta}$ being a consistent estimator that, under H_0, is (asymptotically) normally distributed and A its (asymptotic) covariance-variance matrix computed at $\widehat{\theta}$. These two statistics have many important properties and are widely used in actual problems. Suppose that $X = (X_1, \ldots, X_n)$ is an independent and identically distributed (iid) random sample, under some regular conditions on $L(\theta, X)$ and when Θ_0 is a smooth (semi)algebraic manifold, it is well known that, under H_0, $T(X) = -2\log(\lambda(X))$ converges in distribution to a chisquare distribution with s degrees of freedom (from now on, it is denoted just by χ_s^2), where $s = \dim(\Theta) - \dim(\Theta_0)$ is the co-dimension of Θ_0. The asymptotic distribution of $W(X)$ is a chisquare with rank-of-C degrees of freedom, which is the very same of the likelihood ratio statistics for linear general null hypotheses. We can also mention the Score test statistics that, under appropriated conditions, has asymptotically the same distribution as the two previous statistics. That is, different p-values can be computed for the same problem of hypothesis testing by using different procedures. In this paper we shall only use the procedure based on the (generalized) likelihood ratio statistics, since it has optimal asymptotic properties (see Bahadur and Raghavachari, 2009). From now on, "asymptotic p-values" means p-values computed by using the asymptotic distribution of the test statistic.

Sometimes practitioners have to test a complicated hypothesis H_{01}. By reasons of easiness of computations, instead of testing H_{01}, they may think of testing another auxiliar hypothesis H_{02} such that if H_{02} is false then H_{01} is also false. This procedure is used routinely in medicine and health fields in general, e.g., in genetic studies one of the interests is to test genotype frequencies between two

groups. In this example, we know that H_{01} : "homogeneity of genotype frequencies between the groups" implies H_{02} : "homogeneity of allelic frequencies between the groups". By using mild logical requirements (the consequence condition), if we find evidence against H_{02} we expect to claim evidence against H_{01}. However, as it is widely known, p-values do not follow this logical reasoning.

Let $H_{01} : \theta \in \Theta_{01}$ and $H_{02} : \theta \in \Theta_{02}$ be two null hypotheses such that $\Theta_{01} \subseteq \Theta_{02}$, i.e., H_{01} is nested within H_{02}. It is expected by the logical reasoning to find more evidence against H_{01} than H_{02} for the same observed data $X = x$. In other words, if p_i is the asymptotic p-value computed under H_{0i}, for $i = 1, 2$, respectively, then we expect to observe $p_1 < p_2$. However, if dimensions of the spaces described in these nested hypotheses are different, then their respective asymptotic p-values will be computed under different metrics and therefore inverted conclusions may eventually occur, i.e., more disagreement with H_{02} than H_{01} (i.e., $p_2 < p_1$). That is to say, for a given data and a preassigned α, it may happen $p_1 > \alpha$ and $p_2 < \alpha$. One, therefore, may confront at the same time with "evidence" to reject H_{02} and without "evidence" to reject H_{01}. Of course, the problem here is not with the approximation for the p-values, computed by using limiting reference distribution, the problem also happens with the exact ones. The example below shows the above considerations for exact p-values in a multiparametric scenario.

Example 1.1. *Consider an independent and identically distributed random sample $X = (X_1, \ldots, X_n)$ where $X_1 \sim N_2(\mu, I)$ with $\mu = (\mu_1, \mu_2)^\top$ and I is a (2×2) identity matrix. The full parametric space is $\Theta = \{(\mu_1, \mu_2) : \ \mu_1, \mu_2 \in \mathbb{R}\} = \mathbb{R}^2$. For this example we consider two particular hypotheses. Firstly, suppose that we want to test $H_{01} : \theta \in \Theta_{01}$, where $\Theta_{01} = \{(0, 0)\}$, then the likelihood ratio statistic is*

$$\lambda_1(X) = \frac{\sup_{\theta \in \Theta_{01}} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta, X)} = \exp\left(-\frac{n}{2}\bar{X}^\top \bar{X}\right),$$

where $\bar{X}$ is the sample mean. Taking $T_1(X) = -2\log(\lambda_1(X))$ we know that, under H_{01}, $T_1 \sim \chi_2^2$. Secondly, suppose that the null hypothesis is $H_{02} : \theta \in \Theta_{02}$, where $\Theta_{02} \equiv \{(\mu_1, \mu_2) : \ \mu_1 = \mu_2, \ \mu_1, \mu_2 \in \mathbb{R}\}$, the likelihood ratio statistic is

$$\lambda_2(X) = \frac{\sup_{\theta \in \Theta_{02}} L(\theta; X)}{\sup_{\theta \in \Theta} L(\theta, X)} = \exp\left(-\frac{n}{2}\bar{X}^\top(I - \frac{1}{2}ll^\top)\bar{X}\right),$$

where $l = (1, 1)^\top$. Taking $T_2(X) = -2\log(\lambda_2(X))$ it is possible to show that, under H_{02}, $T_2 \sim \chi_1^2$. Notice that, in this example, the Wald statistics for these two null hypotheses H_{01} and H_{02} are equal to T_1 and T_2, respectively. Assume that the sample size is $n = 100$ and the observed sample mean is $\bar{x} = (0.14, -0.16)^\top$, then $T_1(x) = 4.52$ (with p-value $p_1 = 0.10$) and $T_2(x) = 4.5$ (with p-value $p_2 = 0.03$). These p-values showed evidence against $\mu_1 = \mu_2$, but not against $\mu_1 = \mu_2 = 0$.

However, if we reject that $\mu_1 = \mu_2$ we should technically reject that $\mu_1 = \mu_2 = 0$ (using the very same data).

This issue does not happen only with the likelihood ratio statistic, it happens with many other classical test statistics (score and others) that consider how data should behave under H_0. As p-values are just probabilities to find unobserved statistics, at least, as large as the observed ones, the conflicting conclusion presented in the above example is not a logical contradiction of the frequentist method. *This issue happens because a p-value was not designed to be a measure of evidence over subsets of Θ.* We must say that p-values do exactly the job they were defined to do. However, in the practical scientific world, researches use p-values to take decisions and, hence, they eventually may face some problems with consistency of conclusions. P-values must therefore be used with caution when taking decisions about a null hypothesis. The example below presents a data set which produces surprising conclusions for regression models.

Example 1.2. *Consider a linear model: $y = xb + e$, where $b = (b_1, b_2)^\top$ is a vector formed by two regression parameters, $x = (x_1, x_2)$ is an $(n \times 2)$ matrix of covariates and $e \sim N_n(0, I_n)$ with I_n the $n \times n$ identity matrix. It is usual to verify if each component of b is equal to zero and to remove from the model the non-significant parameters. The majority of statistical routines present the p-values for $H_{0i} : b_i = 0$, say p_i, for $i = 1, 2$. However, sometimes both p-values are greater than α and there exists a joint effect that cannot be discarded. As these hypotheses include a more restricted one, $H_{03} : b = 0$, it is of general advice to remove both b_1 and b_2 only if the p-value p_3, computed for H_{03}, is smaller than α (this decision obeys the logical reasoning). We expect to observe more evidence against H_{03} than either H_{01} and H_{02}. In fact, almost always the p-value p_3 is smaller than both p_1 and p_2, as expected. However, as we shall see below, an inversion of conclusions may occur. To see that, let us present the main ingredients. The maximum likelihood estimator of b is $\widehat{b} = (x^\top x)^{-1} x^\top y$ and the likelihood ratio statistics for testing H_{01}, H_{02} and H_{03} are respectively*

$$\lambda_i(y, x) = \exp\left(-\frac{1}{2} y^\top \left(x(x^\top x)^{-1} x^\top - \dot{x}_i (\dot{x}_i^\top \dot{x}_i)^{-1} \dot{x}_i^\top \right) y \right)$$

for $i = 1, 2$ and

$$\lambda_3(y, x) = \exp\left(-\frac{1}{2} y^\top x (x^\top x)^{-1} x^\top y \right),$$

where $\dot{x}_1 = x_2$ and $\dot{x}_2 = x_1$. It can be showed that $T_i(Y, x) = -2\log(\lambda_i(Y, x)) \sim \chi^2_{s_i}$ for all $i = 1, 2, 3$, where, for this example, $s_1 = s_2 = 1$ and $s_3 = 2$. Again, although $T_3 > T_i$, for $i = 1, 2$, the metrics to compute the p-values are different

and odd behavior may arise as we notice in the following data,

y	-1.29	1.09	-0.16	0.44	-0.22	-1.85	0.91	0.54	0.06	0.37
x_1	3.00	8.00	5.00	9.00	10.00	1.00	6.00	9.00	6.00	5.00
x_2	9.00	7.00	7.00	10.00	7.00	8.00	6.00	6.00	3.00	2.00

Here, the three statistics are $t_1 = 4.48$ (with p-value $p_1 = 0.03$), $t_2 = 4.00$ (with p-value $p_2 = 0.045$) and $t_3 = 4.59$ (with p-value $p_3 = 0.10$). For these data, we have problems with the conclusion, since we expected to have much more evidence against H_{03} than H_{01} and H_{02}. Notice that, $p_3/p_1 \sim 3.3$ and $p_3/p_2 \sim 2.2$.

Many other examples for higher dimensions can be built on, but we think that these two instances are sufficient to illustrate the weakness of p-values when it comes to decide acceptance or rejection of specific hypotheses, for other examples we refer the reader to Schervish (1996). Observe that we used the very same procedure to test both hypotheses H_{01} and H_{02} (i.e., likelihood ratio statistics). Some scientists and practitioners would become confused with these results and it would be very difficult explain to them the reason for that. We believe that the development of a true measure of evidence for null hypotheses that does not have these problems might be welcome by the scientific community.

In summary: in usual frequentist significance tests, a general method of computing test statistics can be used (likelihood ratio statistics, Wald-type statistics, Score statistic and so forth). The distribution of the chosen test statistic depends on the null hypothesis and this leads to different metrics in the computation of p-values (this is the major factor that gives the basis for the frequentist interpretations of p-values). As each of these metrics depend on the dimension of the respective null hypotheses, conflicting conclusions may arise for nested hypotheses. In the next section, we present a new measure that can be regarded as a measure of evidence for null hypotheses without committing any logical contradictions.

This paper unfolds as follows. In Section 2 we present a definition of evidence measure and propose a frequentist version of this measure. Some of its properties are presented in Section 3. A connection with the abstract belief calculus is showed in Section 4. Examples are offered in Section 5. Finally, in Section 6 we discuss the main results and present some final remarks.

2   An evidence measure for null hypotheses

In this section we define a very general procedure to compute a measure of evidence for H_0. The concept of evidence was discussed by Good (1983) in a great philosophical detail. We also refer the reader to Royall (1997) and its review Vieland *et al.* (1998) for relevant arguments to develop new methods of measuring

evidence. As in the previous section, $H_0 : \theta \in \Theta_0$ is the null hypothesis, where $\Theta_0 \subseteq \Theta \subseteq \mathbb{R}^k$ is a smooth manifold. Below, we define what we mean by evidence measure.

Definition 2.1. *Let $\mathcal{X} \subseteq \mathbb{R}^n$ be the sample space. A value $q : \mathcal{X} \times \Theta \to [0, 1]$ is a measure (we shall write just $q(\Theta_0) \equiv q(X, \Theta_0)$ for shortness of notation) of evidence of null hypotheses if the following items hold*

1. *$q(\varnothing) = 0$ and $q(\Theta) = 1$,*

2. *For any two null hypotheses $H_{01} : \theta \in \Theta_{01}$ and $H_{02} : \theta \in \Theta_{02}$, such that $\Theta_{01} \subseteq \Theta_{02}$, we must have $q(\Theta_{01}) \leq q(\Theta_{02})$.*

The above definition is the least we would expect from a coherent measure of evidence. Items 1 and 2 of Definition 2.1 describe a plausibility measure (Friedman and Halpern, 1996). Plausibility measures generalize probability measures, since:

> "A plausibility measure associates with a set a plausibility, which is just an element in a partially ordered space" (Friedman and Halpern, 1996)

As showed in the previous section, *p*-values are not even plausibility measures on Θ, since Condition 2 of Definition 2.1 is not satisfied. Therefore, they cannot be regarded as measures of evidence. Bayes factors are also not plausibility measures on Θ, i.e., Condition 2 fails to be held, (see Lavine and Schervish, 1999; Bickel, 2010). In order to find a purely frequentist measure of evidence, without prior distributions neither over Θ nor H_0, we define a likelihood-based confidence region by

Definition 2.2. *A likelihood-based confidence region with level α is*

$$\Lambda_\alpha = \{\theta \in \Theta : 2(\ell(\widehat{\theta}) - \ell(\theta)) \leq F_\alpha\},$$

where $\widehat{\theta} = \arg\sup_{\theta \in \Theta} \ell(\theta)$ is the maximum likelihood estimator, ℓ is the log-likelihood function and F_α is an $1 - \alpha$ quantile computed from a cumulative distribution function F, i.e., $F(F_\alpha) = 1 - \alpha$. Here, F is (an approximation for) the cumulative distribution function of the random variable $2(\ell(\widehat{\theta}) - \ell(\theta_0))$, where θ_0 is the true value.

Consider the following conditions:

C1. $\widehat{\theta}$ is an interior point of Θ,

C2. ℓ is strictly concave.

Below we define an evidence measure for the null hypothesis H_0.

Definition 2.3. *Let Λ_α be the likelihood-based confidence region. An evidence measure for the null hypothesis $H_0 : \theta \in \Theta_0$ is the value*

$$q = q(\Theta_0) \equiv q(X, \Theta_0) = \max\{0, \sup\{\alpha \in (0, 1) : \ \Lambda_\alpha \cap \Theta_0 \neq \varnothing\}\}.$$

We shall call q-value for short.

In words, q is the greatest significance level for which at least one point of Θ_0 lies inside the confidence region for θ. A simple way of computing a q-value is to build high-confidence regions for θ that includes at least one point of Θ_0 and gradually decreases the confidence until the $(1-q) \times 100\%$ confidence region border intercepts just the last point(s) of Θ_0. The value q is such that the $(1-q+\delta) \times 100\%$ confidence region does not include any points of Θ_0, for any $\delta > 0$. Figure 1 illustrates some confidence regions Λ_α for $\theta = (\theta_1, \theta_2)$ considering different values of α. The dotted line is Λ_{q_1}, where q_1 is the q-value for testing $H_{01} : \theta_1 = 0$. The dot-dashed line is Λ_{q_2}, where q_2 is the q-value for testing $H_{02} : \theta_2 = 0$. The dashed line is Λ_{q_3}, where q_3 is the q-value for testing $H_{03} : \theta_1^2 + \theta_2^2 = 1$.

Observe that a large value of $q(\Theta_0)$ indicates that there exists at least one point in Θ_0 that is near the maximum likelihood estimator, that is, data are not discrediting the null hypothesis H_0. Otherwise, a small value of $q(\Theta_0)$ means that all points of Θ_0 are far from the maximum likelihood estimator, that is, data are discrediting the null hypothesis. The metric that says what is near or far from $\widehat{\theta}$ is the (asymptotic) distribution of $2(\ell(\widehat{\theta}) - \ell(\theta))$. These statements are readily seen by drawing confidence regions (or intervals) with different confidence levels, see Figure 1.

Bickel (2010) developed a method based on the law of likelihood to quantify the weight of evidence for one hypothesis over another. Here, we proposed a classical possibilistic measure over Θ based on likelihood-based confidence regions, see Definition 2.3. Although these approaches are based on similar concepts, they capture different values from the data (we do not investigate this further in the present paper). Another evidence measure that is a Bayesian competitor is the FBST (Full Bayesian Significance Test) proposed originally by Pereira and Stern (1999). See also an invariant version under reparametrizations in Madruga *et al.* (2003) and we refer the reader to Pereira *et al.* (2008) for an extensively review of this latter method.

3 Some important properties

In this section we show some important properties of q-values that will be used to connect them with possibilistic measures and the abstract belief calculus (see Section 4). Our first theorem states that item 1 of Definition 2.1 holds for the proposed q-value.

Theorem 3.1. *Let q be a q-value and consider condition C1, then $q(\varnothing) = 0$ and $q(\Theta) = 1$.*

Proof. As $\widehat{\theta} \in \Theta$ (see condition C1), we have that $\{\alpha \in (0,1) : \Lambda_\alpha \cap \Theta \neq \varnothing\} = (0,1)$ and then $q(\Theta) = 1$. Also, $\{\alpha \in (0,1) : \Lambda_\alpha \cap \varnothing \neq \varnothing\} = \varnothing$, then $q(\varnothing) = \max\{0, \sup(\varnothing)\} = 0$. $\qquad\square$

The following theorem completes the requirement for the q-value to be a measure of evidence.

Theorem 3.2. *(Nested hypotheses) For a fixed data $X = x$, let $H_{01} : \theta \in \Theta_{01}$ and $H_{02} : \theta \in \Theta_{02}$ be two null hypotheses such that $\Theta_{01} \subseteq \Theta_{02}$. Then, $q(\Theta_{01}) \leq q(\Theta_{02})$, where $q(\Theta_{01})$ and $q(\Theta_{02})$ are evidence measures for H_{01} and H_{02}, respectively.*

Proof. Observe that if $\Theta_{01} \subseteq \Theta_{02}$, then $\Lambda_\alpha \cap \Theta_{01} \subseteq \Lambda_\alpha \cap \Theta_{02}$ and $\{\alpha \in (0,1) : \Lambda_\alpha \cap \Theta_{01} \neq \varnothing\} \subseteq \{\alpha \in (0,1) : \Lambda_\alpha \cap \Theta_{02} \neq \varnothing\}$. We conclude that $q(\Theta_{01}) \leq q(\Theta_{02})$ for all $\Theta_{01} \subseteq \Theta_{02} \subseteq \Theta$. $\qquad\square$

Other important feature of our proposal is its invariance under reparametrizations. As likelihood-based confidence regions are invariant under reparametrizations, the q-value is also invariant. Based on Theorem 3.2 we can establish now an interesting result which is related to the Burden of Proof, namely, the evidence in favour of a composite hypothesis is the most favourable evidence in favour of its terms (Stern, 2003).

Lemma 3.1. *(Most Favourable Interpretation) Let I be a countable or uncountable real subset, assume that $\Theta_0 = \bigcup_{i \in I} \Theta_{0i}$ is nonempty, then $q(\Theta_0) = \sup_{i \in I}\{q(\Theta_{0i})\}$.*

Proof. By Theorem 3.2, we know that $q(\Theta_0) \geq q(\Theta_{0i})$ for all $i \in I$, then $q(\Theta_0) \geq \sup_{i \in I}\{q(\Theta_{0i})\}$. To prove this lemma, we must show that $q(\Theta_0) \leq \sup_{i \in I}\{q(\Theta_{0i})\}$.

Define $A(B) = \{\alpha \in (0,1) : \Lambda_\alpha \cap B \neq \varnothing\}$ and note that $A(\Theta_0) \subseteq \bigcup_{i \in I} A(\Theta_{0i})$. Therefore, $\sup(A(\Theta_0)) \leq \sup\left(\bigcup_{i \in I} A(\Theta_{0i})\right) \Rightarrow \sup(A(\Theta_0)) \leq \sup_{i \in I}\{\sup(A(\Theta_{0i}))\}$ and $q(\Theta_0) = \sup_{i \in I}\{q(\Theta_{0i})\}$. $\qquad\square$

Lemma 3.1 states that q-values are possibilistic measures on Θ (Dubois and Prade, 1990; Dubois, 2006). Next lemma presents an important result (for strictly concave log-likelihood functions), which allows us to connect q-values with p-values.

Lemma 3.2. *Assume valid conditions C1 and C2. For a nonempty Θ_0 and F strictly increasing, the q-value can alternatively be defined as*

$$q(\Theta_0) = 1 - F\big(2(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0))\big), \tag{1}$$

where $\widehat{\theta}_0 = \arg\sup_{\theta \in \Theta_0} \ell(\theta)$.

Proof. If Θ_0 is nonempty, there exists $\alpha \in (0,1)$ such that $\Lambda_\alpha \cap \Theta_0 \neq \varnothing$ and the q-value is just $q = \sup\{\alpha \in (0,1) : \Lambda_\alpha \cap \Theta_0 \neq \varnothing\}$. Notice that, as F is strictly increasing, we have that

$$\begin{aligned}
\Lambda_\alpha \cap \Theta_0 &\equiv \{\theta \in \Theta_0 : 2(\ell(\widehat{\theta}) - \ell(\theta)) \leq F_\alpha\} \\
&\equiv \{\theta \in \Theta_0 : F\big(2(\ell(\widehat{\theta}) - \ell(\theta))\big) \leq 1 - \alpha\} \\
&\equiv \{\theta \in \Theta_0 : 1 - F\big(2(\ell(\widehat{\theta}) - \ell(\theta))\big) \geq \alpha\}.
\end{aligned}$$

As ℓ is strictly concave, then for all $0 < \alpha \leq q \leq 1$ we have $\Lambda_q \cap \Theta_0 \subseteq \Lambda_\alpha \cap \Theta_0$ and

$$\sup\{\alpha \in (0,1) : \Lambda_\alpha \cap \Theta_0 \neq \varnothing\} = \sup_{\theta \in \Theta_0} \{1 - F\big(2(\ell(\widehat{\theta}) - \ell(\theta))\big)\}$$

and

$$\sup_{\theta \in \Theta_0} \{1 - F\big(2(\ell(\widehat{\theta}) - \ell(\theta))\big)\} = 1 - F\big(2(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0)),$$

where $\widehat{\theta}_0 = \arg\sup_{\theta \in \Theta_0} \ell(\theta)$. □

The value of $\widehat{\theta}_0$ can be seen as the point of Θ_0 which is in the border of $(1 - q) \times 100\%$ confidence region for θ. Notice that if F is a non-decreasing function than we just can claim that

$$\{\theta \in \Theta_0 : 2(\ell(\widehat{\theta}) - \ell(\theta)) \leq F_\alpha\} \subseteq \{\theta \in \Theta_0 : F\big(2(\ell(\widehat{\theta}) - \ell(\theta))\big) \leq 1 - \alpha\},$$

the converse inclusion may not be valid. In general, F can be approximated to a quisquare distribution with k degrees of freedom, where k is the dimension of Θ (this is a strictly increasing function). Based upon this alternative version we can directly compare q-values with p-values. In addition, it is possible to derive the distribution of q.

Notice that a p-value would be computed by $p = 1 - F_{H_0}\big(2(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0))\big)$, where F_{H_0} is the (asymptotic) distribution of $2(\ell(\widehat{\theta}) - \ell(\widehat{\theta}_0))$ that depends on H_0. On the other hand, the distribution F is the (asymptotic) distribution of $2(\ell(\widehat{\theta}) - \ell(\theta_0))$ which does not depend on H_0, where θ_0 is the true value. Now, by Equation (1), we find a duality between q-values and p-values which is self-evident from Lemma 3.2.

Lemma 3.3. *Assume valid conditions C1 and C2. If F and F_{H_0} are strictly increasing functions, then the following equalities hold*

$$p = 1 - F_{H_0}(F^{-1}(1-q)) \quad and \quad q = 1 - F(F_{H_0}^{-1}(1-p)). \tag{2}$$

Note that, under general regular conditions on the likelihood function and considering that Θ_0 is a smooth semi-algebraic subset of Θ, all asymptotics for the q-value can be derived by using the duality relation presented above in Equation 2. In usual frequentist significance tests, the error probability of type one characterizes the proportion of cases in which a null hypothesis H_0 would be rejected when it is true in a hypothetical long-run of repeated sampling. On the one hand, as a p-value usually has uniform distribution under H_0, the probability to obtain a p-value smaller than α is α. On the other hand, we can only guarantee uniform distribution for the evidence value q under the simple null hypothesis $\Theta_0 = \{\theta_0\}$, which specifies all parameters involved in the model. Since, it can be readily seen that if $\Theta_0 = \{\theta_0\}$, then $F \equiv F_{H_0}$ (an asymptotic quisquare with k degrees of freedom) and therefore $q \sim U(0,1)$ (at least asymptotically). However, if Θ_0 has dimension lesser than k, e.g., under curvature of parameters, the distribution F_{H_0} would differ from F. Notice that the threshold value α adopted for p-values is not valid for q-values, the actual threshold for q-values should be computed using relation (2), i.e., $\alpha_{H_0} = 1 - F(F_{H_0}^{-1}(1-\alpha))$ would be the new cut-off. Of course, if the decision was based on this actual threshold, the same logical contradictions would arise. We remark that, we do not just change the problem to another level, indeed q-values are logical consistent measures of evidence and critical values may be derived in the light of loss functions or other procedures. This cannot be done without committing logical contradictions for p-values, since they are not logically consistent measures of evidence for null hypotheses (see next section).

Pereira *et al.* (2008) left a challenge to the reader, namely, to obtain the one-to-one relationship between the evidence value computed via FBST (e-value) and p-values. Diniz *et al.* (2012) showed that asymptotically the answer is given in Lemma 3.3 replacing the q-value with e-value, therefore, q-values and e-values are asymptotically equivalent.

4   Q-value as an objective state of belief

In this section we analyze the definition of q-values under the light of Abstract Belief Calculus (ABC) proposed by Darwiche and Ginsberg (1992). ABC is a symbolic generalization of probabilities. Probability is a function defined over a family of subsets (known as σ-field) of a main nonempty set Ω to the interval $[0,1]$. The additivity is the main characteristic of this function of subsets, that is, if A

and B are disjoint measurable subsets, then the probability of the union $A \cup B$ is the sum of their respective probabilities. Basically, all theorems of probability calculus require this additive property. Cox (1946) derived this sum rule from a set of more fundamental axioms (see also Jaynes, 1957; Aczel, 2004), which tries to consider the following assertions

> "... the less likely is an event to occur the more likely it is not to occur. The occurrence of both of two events will not be more likely and will generally be less likely than the occurrence of the less likely of the two. But the occurrence of at least one of the events is not less likely and is generally more likely than the occurrence of either" (Cox, 1946).

The quotation above is alleged to be a fundamental part of any coherent reasoning and, based on it, some scholars claim that degrees of belief should be manipulated according to the laws of probability theory (Cox, 1946; Jaynes, 1957; Caticha, 2009). The word "likely" could be replaced by "probable", "possible", "plausible" or any other that represents a measure for our belief or (un)certainty under limited knowledge. Dubois and Prade (2001) said that, under a limited knowledge,

> "one agent that does not believe in a proposition does NOT imply the (s)he believes in its negation"

and also that

> "uncertainty in propositional logic is *ternary* and not binary: either a proposition is believed, or its negation is believed, or neither of them are believed".

Therefore, under a limited knowledge, the claim that "the less possible is an event to occur the more possible it is not to occur" is too restrictive to be of universal applicability in general beliefs.

The Cox's demonstration is made through associativity functional equations (let $G : \mathbb{R}^2 \to \mathbb{R}$ be a real function with two real arguments, then $G(x, G(y, z)) = G(G(x, y), z)$ is the associativity functional equation) that represent the second sentence of the above quotation, the involved function (G) is considered continuous and strictly increasing in both its arguments. However, when this function is non-decreasing we also have a coherent reasoning (see Darwiche and Ginsberg, 1992) and other than additive rules emerge from this functional equation, such as minimum (maximum) as demonstrated by Marichal (2000). Therefore, probability is not the unique coherent way of dealing with uncertainties as usually thought and spread among some scholars. Moreover, Dubois and Prade (2001) clarify that probability theory is not a faithful representation of incomplete knowledge in the sense of classical logic as usual considered.

In our case, the set that we want to define an objective state of belief is the parametric space Θ. Here, the word "objective" means that no prior distributions over Θ are specified. Naturally there exists some level of subjective knowledge in the choice of models, parametric space and so on, these sources of subjectivity will not be discussed further. It is widely known that the frequentist school regards no probability distributions over the subsets of Θ and that probability distributions are only assigned for observable randomized events. Here, we show that the proposed q-value can prescribe an objective state of belief over the subsets of Θ without assigning any prior subjective probability distributions over the subsets of Θ. It is quite obvious that a q-value is not a probability measure, since it is not additive. However, as we shall see in this section, q-values are indeed abstract states of belief.

Abstract belief calculus (ABC) is built considering more basic axioms than regarded in Cox derivation and, as a consequence, the additive property must be generalized to a summation operator $\oplus$ such that the usual sum rule is a simple particular case (they also deal with propositions instead of sets, but here we consider that propositions are represented explicitly as sets). For the sake of completeness, we expose the main components of their theory in what follows. Firstly, let $\mathcal{L}$ be a family of subsets Θ closed under unions, intersections and complements. The ABC starts defining a function $\Phi : \mathcal{L} \to \mathcal{S}$ called support function, and for each $A \in \mathcal{L}$, $\Phi(A)$ is called the support value of A. Then, in order to define the properties of this support function, a partial support structure $\langle \mathcal{S}, \oplus \rangle$ is defined such that the summation support $\oplus : \mathcal{S} \times \mathcal{S} \to \mathcal{S}$ satisfies the following properties:

- Symmetry: $a \oplus b = b \oplus a$ for every $a, b \in \mathcal{S}$

- Associativity: $(a \oplus b) \oplus c = a \oplus (b \oplus c)$, for every $a, b, c \in \mathcal{S}$

- Convexity: For every $a, b, c \in \mathcal{S}$, such that $(a \oplus b) \oplus c = a$, then also $a \oplus b = a$

- Zero element: There exists a unique $\mathbf{0} \in \mathcal{S}$ such that $a \oplus \mathbf{0} = a$ for all $a \in \mathcal{S}$

- Unit element: There exists a unique element $\mathbf{1} \in \mathcal{S}$, where $\mathbf{1} \neq \mathbf{0}$, such that for each $a \in \mathcal{S}$, there exists $b \in \mathcal{S}$ such that $a \oplus b = \mathbf{1}$

Then, the properties of Φ are the following

1. For $A, B \in \mathcal{L}$, such that $A \subseteq B$ and $B \subseteq A$, then $\Phi(A) = \Phi(B)$

2. For $A, B \in \mathcal{L}$, such that $A \cap B = \emptyset$, then $\Phi(A \cup B) = \Phi(A) \oplus \Phi(B)$

3. For $A, B, C \in \mathcal{L}$, such that $A \subseteq B \subseteq C$ and $\Phi(A) = \Phi(C)$, then $\Phi(A) = \Phi(B)$

4. $\Phi(\varnothing) = \mathbf{0}$ and $\Phi(\Theta) = \mathbf{1}$

Backing to our proposal and taking $\mathcal{S} = [0,1]$, $\oplus \equiv$ sup, $\mathbf{0} = 0$, $\mathbf{1} = 1$ and $\Phi \equiv q$ we see, by Theorems 3.1 and 3.2, that the q-value satisfies all the properties above. Therefore, for our problem, we identify that our q-value is acting precisely as a support function Φ in the ABC formalism. It is noteworthy that the support value of A does not determine the support value of A^c (the complement of A). This determination happens in the probability calculus since, when $\mathcal{S} = [0,1]$ and $\oplus \equiv +$, the probabilities of sets that form a partition of the total space must sum up to one. Then, for abstract support functions, Darwiche and Ginsberg (1992) defined the degree of belief function $\ddot{\Phi} : \mathcal{L} \to \mathcal{S} \times \mathcal{S}$, such that $\ddot{\Phi}(A) = \langle \Phi(A), \Phi(A^c) \rangle$. The value of $\ddot{\Phi}(A)$ is said to be the degree of belief of A. Naturally, for the probability calculus this function is vacuous, since if $\mathcal{S} = [0,1]$, $\oplus \equiv +$ and Φ is a probability function then for any measurable subset A, $\ddot{\Phi}(A) = \langle \Phi(A), 1 - \Phi(A) \rangle$.

Now, let $a, b \in \mathcal{S}$, then if there exists $c \in \mathcal{S}$ such that $a \oplus c = b$ we say that the support value a is no greater than the support value b and we use the notation $a \preceq_\oplus b$, the symbol $\preceq_\oplus$ is called as support order. Darwiche and Ginsberg (1992) showed that $\preceq_\oplus$ is a partial order under which $\mathbf{0}$ is minimal and $\mathbf{1}$ is maximal. Also, let $\ddot{\Phi}(A) = \langle a_1, a_2 \rangle$ and $\ddot{\Phi}(B) = \langle b_1, b_2 \rangle$ be degrees of beliefs of A and B, respectively. We say that the degree of belief of A is no greater than the degree of belief of B if $a_1 \preceq_\oplus b_1$ and $b_2 \preceq_\oplus b_1$, this is represented by $\ddot{\Phi}(A) \sqsubseteq_\oplus \ddot{\Phi}(B)$. Darwiche and Ginsberg (1992) also showed that $\sqsubseteq_\oplus$ is a partial order under which $\langle \mathbf{0}, \mathbf{1} \rangle$ is minimal and $\langle \mathbf{1}, \mathbf{0} \rangle$ is maximal. In this coherent framework, $\ddot{\Phi}(A) = \langle \mathbf{1}, \mathbf{1} \rangle$ is fully possible without having any contradictions (of course that in probability measures this cannot happen). The authors clarified this in the quotation below:

> "A sentence is rejected precisely when it is supported to degree $\mathbf{0}$. And a sentence is accepted only if it is supported to degree $\mathbf{1}$. But if the sentence is supported to degree $\mathbf{1}$, it is not necessarily accepted. For example, when degrees of support are $\mathcal{S} = \{possible, impossible\}$, a sentence and its negation could be *possible*. Here, both the sentence and its negation are supported to degree *possible*, but neither is accepted."

As our q-value is a support function for the subsets of Θ, we may used this measure to state degrees of belief for the subsets of Θ. Notice that we cannot do this with the usual concept of p-value by the following. If $A \subset B \subset \Theta$, then A e $A^c \cap B$ are disjoint sets, therefore as $B = A \cup (A^c \cap B)$ we have by Property 2 that $\Phi(B) = \Phi(A) \oplus \Phi(A^c \cap B)$ and then $\Phi(A) \preceq_\oplus \Phi(B)$, that is, if A is a subset of B the support value $\Phi(A)$ is no greater than the support value $\Phi(B)$. By our Examples 1.1 and 1.2 we see that p-values do not satisfy this requirement. As aforementioned, q-values are not probabilities measures on the subsets of Θ

and this is not a weakness as some may argue (for subsets of Θ with dimensions smaller than Θ, the best measure of evidence that probability measures can provide is zero). We remark that p-values and the Bayesian e-values (Pereira and Stern, 1999) are not probability measures on Θ, moreover, usually defined p-values cannot even be included in the ABC formalism to establish objective states of belief on the subsets of Θ. Hence, q-values can be used to fill this gap.

Consider the null hypothesis $H_0 : \theta \in \Theta_0$, by Condition C1, we know that $\widehat{\theta} \in \Theta_0$ or $\widehat{\theta} \in \Theta_0^c$, then $q(\Theta_0) = 1$ or $q(\Theta^c) = 1$. As the q-value is a measure of support and $\langle 0, 1 \rangle$ is minimal and $\langle 1, 0 \rangle$ is maximal, we can readily reject H provided that $\ddot{\Phi}(\Theta_0) = \langle q(\Theta_0), q(\Theta_0^c) \rangle = \langle 0, 1 \rangle$ and readily accept H provided that $\ddot{\Phi}(\Theta_0) = \langle q(\Theta_0), q(\Theta_0^c) \rangle = \langle 1, 0 \rangle$. In these two cases we have complete knowledge. When $\ddot{\Phi}(\Theta_0) = \langle q(\Theta_0), q(\Theta_0^c) \rangle = \langle 1, 1 \rangle$ we have complete ignorance regarding this specific hypothesis and cannot either accept or reject, then we must perform other experiment for gathering more information. In general, we have intermediate states of knowledge, namely: (1) $\ddot{\Phi}(\Theta_0) = \langle a, 1 \rangle$ or (2) $\ddot{\Phi}(\Theta_0) = \langle 1, b \rangle$, where $a, b \in [0, 1]$. In the first case, we say that there is evidence against H_0 if a is sufficiently small (i.e., $a < a_c$ for some critical value a_c) and we say that the decision is unknown whenever a is not sufficiently small. In the second case, we say that there is evidence in favor of H_0 if b is sufficiently small (i.e., $b < b_c$ for some critical value b_c) and we say that the decision is unknown whenever b is not sufficient small. The problem now is to define what is "sufficiently small" to perform a decision. The decision rule may be derived through loss functions or other procedure, we will study this issue in future works. Whatever the chosen procedure, it should respect the minimal ($\ddot{\Phi}(\Theta_0) = \langle 0, 1 \rangle$), maximal ($\ddot{\Phi}(\Theta_0) = \langle 1, 0 \rangle$) and inconclusive ($\ddot{\Phi}(\Theta_0) = \langle 1, 1 \rangle$) features of possibilistic measures (we stress that p-values have a similar behaviour, since they are almost possibilistic measures).

Notice that if the null and alternative hypotheses ($H_0 : \theta \in \Theta_0$ and $H_1 : \theta \in \Theta_0^c$, respectively) are fixed before observing the data, it is possible to obtain a maximum likelihood estimate that lies in the null parametric space, $\hat{\theta} \in \Theta_0$. For instance, testing the mean of a normal distribution $H_0 : \mu \leq 0$ against $H_1 : \mu > 0$, we may eventually observe a negative sample average, $\bar{X} \leq 0$, then the respective p-value computed by using likelihood ratio statistics will equal one. This happens in general provided $\hat{\theta} \in \Theta_0$. In practice, practitioners formulate a hypothesis suggested by the data to avoid such feature (i.e., in the mean testing of the normal distribution: if $\bar{X} > 0$, then "$H_0 : \mu \leq 0$ against $H_1 : \mu > 0$"; if $\bar{X} < 0$, then "$H_0 : \mu \geq 0$ against $H_1 : \mu < 0$"; the null and alternative hypotheses are interchanged with modifications for the null parametric space being closed). If the maximum likelihood estimate lies in the border of the null parametric space, then the respective p-value will generally equal one for both cases. For instance, in the normal example, if $\bar{X} = 0$ the respective p-values will be one for either both

cases: "$H_0 : \mu \leq 0$ against $H_1 : \mu > 0$" and "$H_0 : \mu \geq 0$ against $H_1 : \mu < 0$". As q-values are support measures, we do not need to interchange the null and alternative hypotheses, all we have to do is to analyze the hypothesis according to the rules of possibilistic measures stated in the previous paragraph.

5   Examples

In this section we apply our proposal to the Examples 1.1 and 1.2 and we also consider an example for the Hardy-Weinberg equilibrium hypothesis.

Example 5.1. *Consider Example 1.1, after a straightforward computation we find*

$$2(\ell(\widehat{\theta}) - \ell(\theta)) = n(\bar{X} - \mu)^\top (\bar{X} - \mu),$$

where $\bar{X} = (\bar{X}_1, \bar{X}_2)^\top$ and $\mu = (\mu_1, \mu_2)^\top$ and F_α is the α quantile from a quisquare distribution with two degrees of freedom. Then, q-values for $H_{01} : \mu_1 = \mu_2 = 0$ and $H_{02} : \mu_1 = \mu_2$ are respectively $q_1 = 0.104$ and $q_2 = 0.105$. Note that the curve $\mu_1 = \mu_2$ intercepts Λ_{q_2} at $(-0.01, -0.01)$. As expected for this case, $q_1 < q_2$, since $\Theta_{01} \subset \Theta_{02}$, in addition, q_1 and q_2 are near each other because the variables are independent. If the variables were correlated, those q-values would differ drastically (being q_2 always greater than q_1).

Example 5.2. *Consider Example 1.2, the maximum likelihood estimates for b_1 and b_2 are respectively $\widehat{b}_1 = 0.1966$ and $\widehat{b}_2 = -0.1821$. Here, we find that*

$$2(\ell(\widehat{\theta}) - \ell(\theta)) = (\widehat{b} - b)^\top (x^\top x)(\widehat{b} - b).$$

Then, F_α is the α quantile from a quisquare distribution with two degrees of freedom and the q-values for $H_{01} : b_1 = 0$, $H_{02} : b_2 = 0$ and $H_{03} : (b_1, b_2) = (0, 0)$ are respectively $q_1 = 0.107$, $q_2 = 0.135$ and $q_3 = 0.101$. Therefore, as expected by the logical reasoning $q_1 > q_3$ and $q_2 > q_3$.

We also compare our results with the FBST approach considering a trinomial distribution and the Hardy-Weinberg equilibrium hypothesis.

Example 5.3. *Consider that we observe a vector of three values x_1, x_2, x_3, in which the likelihood function is proportional to $\theta_1^{x_1} \theta_2^{x_2} \theta_3^{x_3}$, where $x_1 + x_2 + x_3 = n$ and the parametric space is $\Theta = \{\theta \in (0, 1)^3 : \theta_1 + \theta_2 + \theta_3 = 1\}$. Here, we use the same settings described in Section 4.3 by Pereira and Stern (1999), that is, the null hypothesis is $\Theta_0 = \left\{\theta \in \Theta : \theta_3 = (1 - \sqrt{\theta_1})^2\right\}$ and $n = 20$.*

Table 6 presents the q-values for all values of x_1 and x_3. The last two columns were taken from Table 2 of Pereira and Stern (1999). It should be said that we

computed the q-values by using Definition 2.3 instead of Relation (2), because the p-values were presented with two decimal places in Pereira and Stern (1999) and this can induce distorted q-values. As it was seen, our proposal yields similar results to the FBST approach.

6   Discussion and final remarks

Berger and Sellke (1987) compare p-values with posterior probabilities (by using objective prior distributions) and find differences by an order of magnitude (when testing a normal mean, data may produce a p-value of 0.05 and posterior probability of the null hypothesis of at least 0.30). As opposed to posterior distributions, p-values do not hold the requirement of evidence measures, but one can also conclude that posterior probabilities cannot be used to reflect probabilities in a hypothetical long-run of repeated sampling. Also, posterior probabilities cannot provide a measure of evidence different from zero under sharp hypotheses (when the dimension of the null parametric space is smaller than the full parametric space). A Bayesian procedure that provides a positive evidence measure under sharp hypotheses is the FBST. In this context, q-values are directly comparable with the evidence measures of the FBST approach. This latter procedure needs numerical integrations and maximizations, which may be difficult to be attained for high dimensional problems. As we studied in previous sections, our procedure produces similar results to the FBST and can be readily used as a classical alternative (when the user does not want to specify prior distributions). Moreover, if one has a p-value computed via likelihood ratio statistic, then Relation (2) may be applied to compute the respective q-value without any further computational procedures (maximizations and integrations) and, also, this relation allows to derive the q-value distribution (if desired). Also, we should mention that q-values, when computed using the asymptotic distribution of $2(\ell(\widehat{\theta}) - \ell(\theta_0))$, do respect the famous likelihood principle, but we should say that it is not the main concern here, it is just a property of our approach. Naturally, if the exact distribution of $2(\ell(\widehat{\theta}) - \ell(\theta_0))$ is adopted, the likelihood principle may be violated.

When the null hypothesis is simple and specifies the full vector of parameters, say $\Theta_0 = \{\theta_0\}$, the proposed q-values are, in general, p-values. Otherwise, q-values cannot be interpreted as p-values, instead, they must be treated as measures of evidence for null hypotheses. As aforementioned, when treated as evidence measures, p-values have some internal undesirable features (in some cases, for nested hypotheses H_{01} and H_{02}, where H_{01} is nested within H_{02}, p-values might give more evidence against H_{02} than H_{01}). On the other hand, p-values respect the repeated sampling principle, that is, in the long-run average actual error of rejecting a true hypothesis is not greater than the reported error. In other words,

as p-values have uniform distribution under the null hypothesis, the frequency of observing p-values smaller than α is α. This is an external desirable aspect, since this allows us to verify model assumptions and sensibility, among many other things. The proposed q-values overcome that internal undesirable aspect of p-values, but the problem now is how to evaluate a critical value to establish a decision rule for a hypothesis based on q-values (this decision should respect the rules of possibilistic measures). If we want to respect the repeated sampling principle, based on Lemma 3.3, we see that a critical value for q depends on H_0. To see that, let α be the chosen critical value for the computed p-value, then it can be "corrected" to $\alpha'_{H_0} = 1 - F(F_{H_0}^{-1}(1 - \alpha))$ for the respective q-value. This threshold value α'_{H_0} will respect the repeated sampling principle if and only if it varies with H_0. If we adopt this "corrected" critical value we will have the same internal undesirable features of p-values. We must rely on other principles to compute the threshold value for our q-value, maybe based on loss functions. These loss functions may incorporate the scientific importance of a hypothesis to elaborate a reasonable critical value (this issue will be discussed in future work). It is well known that statistical significance is not the same as scientific significance, for a further discussion we refer the reader to Cox (1977). Naturally, we could also employ loss functions on p-values to find a threshold, however, the internal undesirable features of p-values will certainly bring problems to implement this without any logical conflicts.

There are many open issues that need more attention regarding q-values. Next we provide a list of open problems that we did not deal with in this article, but will be subject of our future research.

1. To give a rigorous mathematical treatment when the log-likelihood function ℓ is not strictly concave.

2. To derive a computational procedure to find q-values and their distribution for (semi)algebraic subsets Θ_0 and not strictly concave ℓ.

3. To compare theoretical properties of q-values by using other types of confidence regions. Monte Carlo simulations may be required.

4. To compare q-values with evidence values (e-value) computed via FBST (Pereira and Stern, 1999) and other procedures such as the posterior Bayes factor (Aitkin, 1991) in a variety of models by using actual data.

5. To derive a criterion to advise one out of three decisions "acceptance", "rejection" or "undecidable" of a null hypothesis without having any types of conflict.

We end this paper by saying that we are not advocating a replacement of p-values by q-values. Instead, we just recommend q-values as additional measures to assist data analysis.

Acknowledgements

References

Aczel, J. (2004). The Associativity Equation ReRevisited, *AIP Conf. Proc.* **707**, 195–203.

Aitkin, M. (1991). Posterior Bayes Factors, *Journal of the Royal Statistical Society – Series B*, **53**, 111–142.

Bahadur, R.R., Baghavachari, M. (2009). Some asymptotic properties of likelihood ratios on general sample spaces, *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob.*, Vol. 1 (Univ. of Calif. Press, 1972), 129–152.

Berger, J.O., Sellke, T. (1987). Testing a point null hypothesis: The irreconcilability of p values and evidence, *Journal of the American Statistical Association*, **82**, 112–122.

Bickel, D.R. (2008). The strength of statistical evidence for composite hypotheses: Inference to the Best Explanation. Technical Report, Ottawa Institute of Systems Biology, COBRA Preprint Series, Article 71, available at http://biostats.bepress.com/cobra/art71/

Birkes, D. (1990). Generalized Likelihood Ratio Tests and Uniformly Most Powerful Tests, *The American Statistician*, **44**, 163–166.

Caticha, A. (2009). Quantifying Rational Belief, arxiv.org/abs/0908.3212

Cox, D.R. (1977). The role of significant tests (with discussion), *Scandinavian Journal of Statistics*, **4**, 49–70.

Cox, R.T. (1946). Probability, Frequency and Reasonable Expectation, *American Journal of Physics*, **14**, 1–13.

Darwiche, A.Y., Ginsberg, M.L. (1992). A symbolic generalization of probability theory, *AAAI-92, Tenth National Conference on Artificial Intelligence.*

Diniz, M., Pereira, C.A.B., Polpo, A., Stern, J.M., Wechsler, S. (2012). Relationship between Bayesian and Frequentist significance indices, *International Journal for Uncertainty Quantification*, **2**, 161–172.

Dubois, D. (2006). Possibility theory and statistical reasoning, *Computational Statistics & Data Analysis*, **51**, 47–69.

Dubois, D. and Prade, H. (1990). An introduction to possibilistic and fuzzy logics. In G. Shafer and J. Pearl (Eds.), Readings in Uncertain Reasoning, 742–761. San Francisco: Morgan Kaufmann.

Dubois, D. and Prade, H. (2001). Possibility theory, probability theory and multiple-valued logics: A clarification, *Annals of Mathematics and Artificial Intelligence*, **32**, 35–66.

Friedman, N., Halpern, J.Y. (1996). Plausibility measures and default reasoning, *Journal of the ACM*, **48**, 1297–1304.

Grendár, A. (2012). Is the p-value a good measure of evidence? An asymptotic consistency criterion, *Statistics & Probability Letters*, **86**, 1116–1119.

Good, I.J. (1983). *Good thinking: The foundations of probability and its applications*; University of Minnesota Press, 1983; p 332.

Jaynes, E.T. "How does the Brain do Plausible Reasoning", Stanford Univ. Microwave Lab. report 421 (1957)

Marichal J-L. (2000). On the associativity functional equation, *Fuzzy Sets and Systems*, **114**, 381–389.

Mayo, D. (2004). "An Error-Statistical Philosophy of Evidence" in M. Taper and S. Lele (eds.) The Nature of Scientific Evidence: Statistical, Philosophical and Empirical Considerations. Chicago: University of Chicago Press: 79–118 (with discussion).

Mayo D.G., Cox D.R. Frequentist statistics as a theory of inductive inference, 2nd Lehmann Symposium – Optimality IMS Lecture Notes – Mongraphs Series (2006).

Mayo, D., Spanos, A. (2006). Severe Testing as a Basic Concept in a NeymanPearson Philosophy of Induction, *Brit. J. Phil. Sci.*, **57**, 323–357.

Kempthorne, O. (1976). Of what use are tests of significance and tests of hypothesis, *Communications in Statistics – Theory and Methods*, **8**, 763–777.

Lavine, M., Schervish, M.J. (1999). Bayes factors: What they are and what they are not, *The American Statistician*, **53**, 119–122.

Madruga, M., Pereira, C.A.B., Stern, J.M. (2003). Bayesian Evidence Test for Precise Hypotheses, *Journal of Statistical Planning and Inference*, **117**, 185–198.

Pereira, C.A.B. and Stern, J.M. (1999). Evidence and credibility: Full Bayesian significance test for precise hypotheses, *Entropy*, **1**, 99–110.

Pereira, C.A.B., Stern, J.M., Wechsler, S. (2008). Can a significance test be genuinely Bayesian?, *Bayesian Analysis*, **3**, 79–100.

Pereira, C.A.B. and Wechsler, S. (1993). On the concept of P-value, *Brazilian Journal of Probability and Statistics*, **7**, 159–177.

Rice, K. (2010). A Decision-Theoretic Formulation of Fishers Approach to Testing, *The American Statistician*, **64**, 345–349.

Royall, R. (1997). *Statistical Evidence: A Likelihood Paradigm*; Chapman & Hall: London; p 191.

Schervish, M.J. (1996). P Values: What they are and what they are not, *The American Statistician*, **50**, 203–206.

Stern, J.M. (2003). Significance tests, Belief Calculi, and Burden of Proof in Legal and Scientific Discourse, *Frontiers in Artificial Intelligence and Applications, Amsterdan*, **101**, 139–147.

Vieland, V.J.; Hodge, S.E. Book Reviews: Statistical Evidence by R. Royall (1997), *Am. J. Hum. Genet.*, **63**, 283–289.

Wagenmakers, E-J. (2007). A practical solution to the pervasive problems of p values, *Psychonomic Bulletin & Review*, **14**, 779–804.

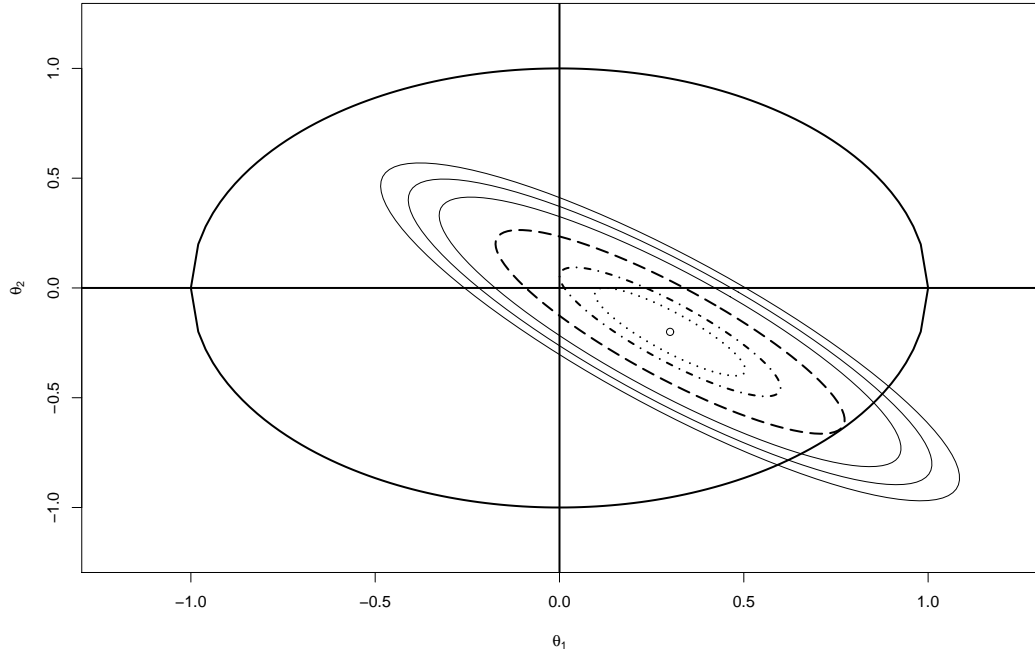Zang, Z. (2009). A Law of Likelihood for Composite Hypotheses, arxiv.org/abs/0901.0463

Figure 1: Borders of confidence regions Λ_α, for different values of α. The dotted line is Λ_{q_1}, where q_1 is the q-value for testing $H_{01} : \theta_1 = 0$. The dot-dashed line is Λ_{q_2}, where q_2 is the q-value for testing $H_{02} : \theta_2 = 0$. The dashed line is Λ_{q_3}, where q_3 is the q-value for testing $H_{03} : \theta_1^2 + \theta_2^2 = 1$.

Table 1: Tests of Hardy-Weinberg equilibrium

x_1	x_3	q-value	e-value (FBST)	p-value
1	2	0.00	0.01	0.00
1	3	0.02	0.01	0.01
1	4	0.04	0.04	0.02
1	5	0.10	0.09	0.04
1	6	0.20	0.18	0.08
1	7	0.33	0.31	0.15
1	8	0.50	0.48	0.26
1	9	0.68	0.66	0.39
1	10	0.84	0.83	0.57
1	11	0.95	0.95	0.77
1	12	1.00	1.00	0.99
1	13	0.96	0.96	0.78
1	14	0.85	0.84	0.55
1	15	0.68	0.66	0.33
1	16	0.48	0.47	0.16
1	17	0.29	0.27	0.05
1	18	0.13	0.12	0.00
5	0	0.01	0.02	0.01
5	1	0.10	0.09	0.04
5	2	0.32	0.29	0.14
5	3	0.63	0.61	0.34
5	4	0.90	0.89	0.65
5	5	1.00	1.00	1.00
5	6	0.91	0.90	0.66
5	7	0.69	0.66	0.39
5	8	0.44	0.40	0.20
5	9	0.24	0.21	0.09
5	10	0.11	0.09	0.04
9	0	0.12	0.21	0.09
9	1	0.68	0.66	0.39
9	2	0.99	0.99	0.91
9	3	0.87	0.86	0.59
9	4	0.53	0.49	0.26
9	5	0.24	0.21	0.09
9	6	0.08	0.06	0.03
9	7	0.02	0.01	0.01